Analysis of Optimal Transport Related Misfit Functions in Seismic Imaging

Yunan Yang(✉) and Björn Engquist

Department of Mathematics, The University of Texas at Austin,
University Station C1200, Austin, TX 78712, USA
yunanyang@math.utexas.edu

Abstract. We analyze different misfit functions for comparing synthetic and observed data in seismic imaging, for example, the Wasserstein metric and the conventional least-squares norm. We revisit the convexity and insensitivity to noise of the Wasserstein metric which demonstrate the robustness of the metric in seismic inversion. Numerical results illustrate that full waveform inversion with quadratic Wasserstein metric can often effectively overcome the risk of local minimum trapping in the optimization part of the algorithm. A mathematical study on Fréchet derivative with respect to the model parameters of the objective functions further illustrates the role of optimal transport maps in this iterative approach. In this context we refer to the objective function as misfit. A realistic numerical example is presented.

Keywords: Full waveform inversion · Optimal transport · Seismic imaging · Optimization · Inverse problem

1 Introduction

Seismic data contains interpretable information about subsurface properties. Imaging predicts the spatial locations as well as properties that are useful in exploration seismology. The inverse method in the imaging predicts more physical properties if a full wave equation is employed instead of an asymptotic far-field approximation to it [9].

This, so called full waveform inversion (FWI) is a data-driven method to obtain high resolution subsurface properties by minimizing the difference or misfit between observed and synthetic seismic waveforms [12]. In the past three decades, the least-squares norm (L^2) has been widely used as a misfit function [10], which is known to suffer from cycle skipping issues (local minimum trapping) and sensitivity to noise [12].

Optimal transport has become a well developed topic in mathematics since it was first proposed by Gaspard Monge in 1781. The idea of using optimal transport for seismic inversion was first proposed in 2014 [3]. A useful tool from the theory of optimal transport, the Wasserstein metric computes the optimal cost of rearranging one distribution into another given a cost function. In computer

F. Nielsen and F. Barbaresco (Eds.): GSI 2017, LNCS 10589, pp. 109–116, 2017.
https://doi.org/10.1007/978-3-319-68445-1_13

science the metric is often called the "Earth Mover's Distance" (EMD). Here we will focus on the quadratic Wasserstein metric (W_2).

In this paper, we briefly review the theory of optimal transport and revisit the convexity and noise insensitivity of W_2 that were proved in [4]. The properties come from the analysis of the objective function. Next, we compare the Fréchet derivative with respect to the model parameters in different misfit functions using the adjoint-state method [8]. Discussions and comparisons between large scale inversion results using W_2 and L^2 metrics illustrate that the W_2 metric is very promising for overcoming the cycle skipping issue in seismic inversion.

2 Theory

2.1 Full Waveform Inversion and the Least Squares Functional

Full waveform inversion is a PDE-constrained optimization problem, minimizing the data misfit $J(f, g)$ by updating the model m, i.e.:

$$m^\star = \underset{m}{\operatorname{argmin}}\ J(f(x_r, t; m), g(x_r, t)), \tag{1}$$

where g is observed data, f is simulated data, x_r are receiver locations. We get the modeled data $f(x, t; m)$ by numerically solving in both the space and time domain [1].

Generalized least squares functional is a weighted sum of the squared errors and hence a generalized version of the standard least-squares misfit function. The formulation is

$$J_1(m) = \sum_r \int |W(f(x_r, t; m)) - W(g(x_r, t))|^2\, dt. \tag{2}$$

In the conventional L^2 misfit, the weighting operator W is the identity I.

The integral wavefields misfit functional [5] is a generalized least squares functional applied on full-waveform inversion (FWI) with weighting operator $W(u) = \int_0^t u(x, \tau) d\tau$. If we define the integral wavefields $U(x, t) = \int_0^t u(x, \tau) d\tau$, then misfit function becomes the ordinary least squares difference between $\int_0^t g(x_r, \tau) d\tau$ and $\int_0^t f(x_r, \tau; m) d\tau$. The integral wavefields still satisfy the original acoustic wave equation with a different source term: $\delta(\boldsymbol{x} - \boldsymbol{x}_s) H(t) * S(t)$, where S is the original source term and $H(t)$ is the Heaviside step function [5]. We will refer this misfit function as H^{-1} norm in this paper.

Normalized Integration Method (NIM) is another generalized least squares functional, with an additional normalization step than integral wavefields misfit functional [6]. The weighting operator is

$$W(u)(x_r, t) = \frac{\int_0^t P(u)(x_r, \tau) d\tau}{\int_0^T P(u)(x_r, \tau) d\tau}, \tag{3}$$

where function P is included to make the data nonnegative. Three common choices are $P_1(u) = |u|$, $P_2(u) = u^2$ and $P_3 = E(u)$, which correspond to the absolute value, the square and the envelop of the signal [6].

2.2 Optimal Transport

Optimal transport is a problem that seeks the minimum cost required to transport mass of one distribution into another given a cost function, e.g. $|x-y|^2$. The mathematical definition of the distance between the distributions $f : X \to \mathbb{R}^+$ and $g : Y \to \mathbb{R}^+$ can then be formulated as

$$W_2^2(f,g) = \inf_{T_{f,g}\in\mathcal{M}} \int_X |x - T_{f,g}(x)|^2 \, f(x)\, dx \tag{4}$$

where $\mathcal{M}$ is the set of all maps $T_{f,g}$ that rearrange the distribution f into g [11].

The Wasserstein metric is an alternative misfit function for FWI to measure the difference between synthetic data f and observed data g. We can compare the data trace by trace and use the Wasserstein metric (W_p) in 1D to measure the misfit. The overall misfit is then

$$J_2(m) = \sum_{r=1}^{R} W_p^p(f(x_r,t;m), g(x_r,t)), \tag{5}$$

where R is the total number of traces. In this paper, we mainly discuss about quadratic Wasserstein metric (W_2) when $p=2$ in (4) and (5).

Here we consider $f_0(t)$ and $g_0(t)$ as synthetic data and observed data from one single trace. After proper scaling with operator P, we get preconditioned data $f = P(f_0)$ and $g = P(g_0)$ which are positive and having total sum one. If we consider they are probability density functions (pdf), then after integrating once, we get the cumulative distribution function (cdf) $F(t)$ and $G(t)$.

If f is continuous we can write the explicit formulation for the 1D Wasserstein metric as:

$$W_2^2(f,g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt = \int_0^T (G^{-1}F(t) - t)^2 f(t) dt. \tag{6}$$

The interesting fact is that W_2 computes the L^2 misfit between F^{-1} and G^{-1} (Fig. 1), while the objective function of NIM measures the L^2 misfit between F and G (Fig. 1). This is identical to the mathematical norm of Sobolev space H^{-1}, $||f-g||_{H^{-1}}^2$, given f and g are nonnegative and sharing equal mass.

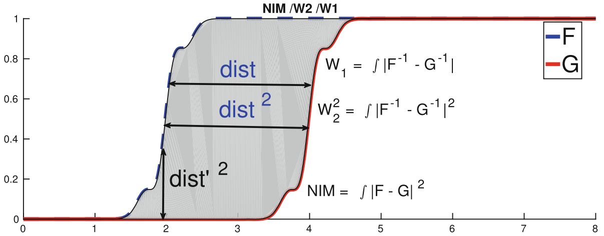
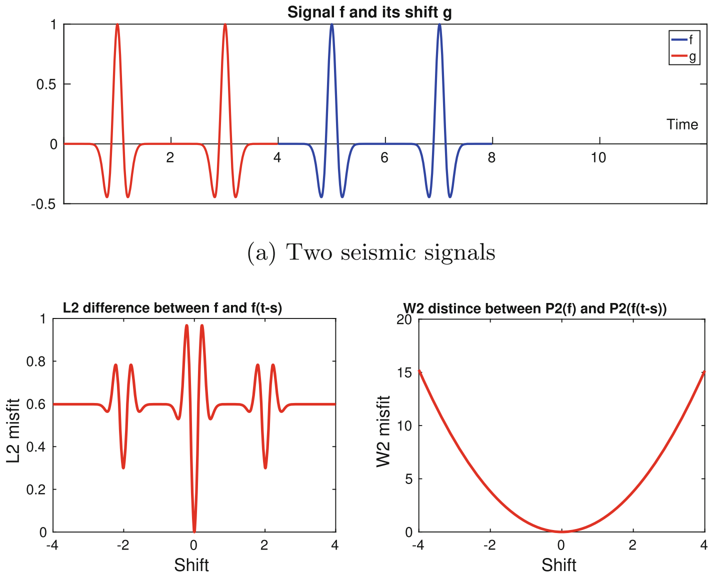


Fig. 1. After data normalization NIM measures $\int(F-G)^2 dt$, while W_2 considers $\int(F^{-1}-G^{-1})^2 dt$ and W_1 considers $\int|F^{-1}-G^{-1}|dt$.

3 Properties

Figure 2a shows two signals f and its shift g, both of which contain two ricker wavelets. Shift of signals are common in seismic data when we have incorrect velocity. We compute the L^2 norm and W_2 norm between f and g, and plot the misfit curves in terms of s in Fig. 2b and c. The L^2 difference between two signals has many local minima and maxima as s changes. It is a clear demonstration of the cycle skipping issue of L^2 norm. The global convexity of Fig. 2c is a motivation to further study the ideal properties of W_2 norm.



(a) Two seismic signals

(b) L^2 misfit

(c) W_2 misfit

Fig. 2. (a) A signal consisting two Ricker wavelets (blue) and its shift (red). (b) L^2 norm between f and g which is a shift of f. (c) W_2 norm between $P_2(f)$ and $P_2(g)$ in terms of different shift s. (Color figure online)

As demonstrated in [4], the squared Wasserstein metric has several properties that make it attractive as a choice of misfit function. One highly desirable feature is its convexity with respect to several parameterizations that occur naturally in seismic waveform inversion [13]. For example, variations in the wave velocity lead to simulations $f(m)$ that are derived from shifts,

$$f(x; s) = g(x + s\eta), \quad \eta \in \mathbb{R}^n, \tag{7}$$

or dilations,

$$f(x; A) = g(Ax), \quad A^T = A,\, A > 0, \tag{8}$$

applied to the observation g. Variations in the strength of a reflecting surface or the focusing of seismic waves can also lead to local rescalings of the form

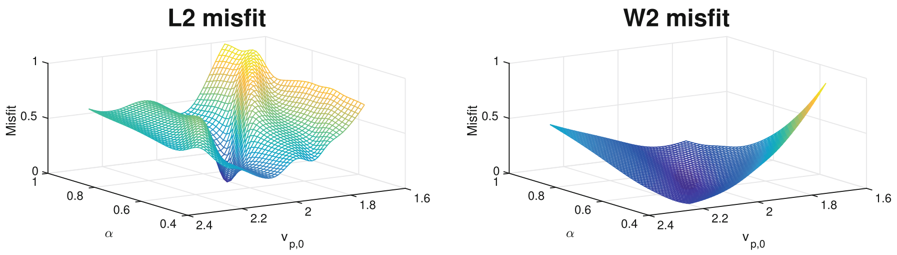
$$f(x;\beta) = \begin{cases} \beta g(x), & x \in E \\ g(x), & x \in \mathbb{R}^n \backslash E. \end{cases} \tag{9}$$

In Theorem 1, f and g are assumed to be nonnegative with identical integrals.

Theorem 1 (Convexity of squared Wasserstein metric [4]). *The squared Wasserstein metric $W_2^2(f(m), g)$ is convex with respect to the model parameters m corresponding to a shift s in (7), the eigenvalues of a dilation matrix A in (8), or the local rescaling parameter β in (9).*

The Fig. 2c numerically exemplifies Theorem 1. Even if the scaling $P(u) = u^2$ perfectly fits the theorem it has turned out not to work well in generating an adjoint source that works well in inversion. The linear scaling, $P(u) = au + b$, on the other hand works very well even if the related misfit lacks strict convexity with respect to shifts. The two-variable example described below and Fig. 3 are based on the linear scaling. It gives the convexity with respect to other variables in velocity than a simple shift in the data.

The example from [7] shows a convexity result in higher dimensional model domain. The model velocity is increasing linearly in depth as $v(x, z) = v_{p,0} + \alpha z$, where $v_{p,0}$ is the starting velocity on the surface, α is vertical gradient and z is depth. The reference for $(v_{p,0}, \alpha)$ is $(2\,\mathrm{km/s},\ 0.7\,\mathrm{s}^{-1})$, and we plot the misfit curves with $\alpha \in [0.4, 1]$ and $v_0 \in [1.75,\ \ 2.25]$ on 41×45 grid in Fig. 3. We observe many local minima and maxima in Fig. 3a. The curve for W_2 (Fig. 3b) is globally convex in model parameters $v_{p,0}$ and α. It demonstrates the capacity of W_2 in mitigating cycle skipping issues.



(a) L^2 (b) W_2

Fig. 3. (a) Conventional L^2 misfit function (b) W_2 misfit function trace-by-trace

Another ideal property of optimal transport is the insensitivity to noise. All seismic data contains either natural or experimental noise. For example, the ocean waves lead to extremely low frequency data in marine acquisition. Wind and cable motions also generate random noise.

The L^2 norm is known to be sensitive to noise since the misfit between clean and noisy data is calculated as the sum of squared noise amplitude at each sampling point. In [4] W_2 norm is proved to be insensitive to mean-zero noise and the property apply for any dimension of the data. This is a natural result from optimal transport theory since the W_2 metric defines a global comparison that not only considers the change in signal intensity but also the phase difference.

Theorem 2 (Insensitivity to noise [4]**).** *Let f_{ns} be f with a piecewise constant additive noise of mean zero uniform distribution. The squared Wasserstein metric $W_2^2(f, f_{ns})$ is of $\mathcal{O}(\frac{1}{N})$ where N is the number of pieces of the additive noise in f_{ns}.*

4 Discussions

Typically we solve the linearized problem iteratively to approximate the solution in FWI. This approach requires the Fréchet derivatives of the misfit function $J(m)$ which is expensive to compute directly. The adjoint-state method [8] provides an efficient way of computing the gradient. This approach requires the Fréchet derivative $\frac{\partial J}{\partial f}$ and two modelings by solving the wave equations. Here we will only discuss about the acoustics wave Eq. (10).

$$m\frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u(x,t) = S(x,t) \tag{10}$$

In the adjoint-state method, we first forward propagate the source wavelet with zero initial conditions. The simulated data f is the source wavefield u recorded on the boundary. Next we back propagate the Fréchet derivative $\frac{\partial J}{\partial f}$ as the source with zero final conditions and get the receiver wavefield v.

With both the forward wavefield u and backward wavefield v, the Fréchet derivative of m becomes

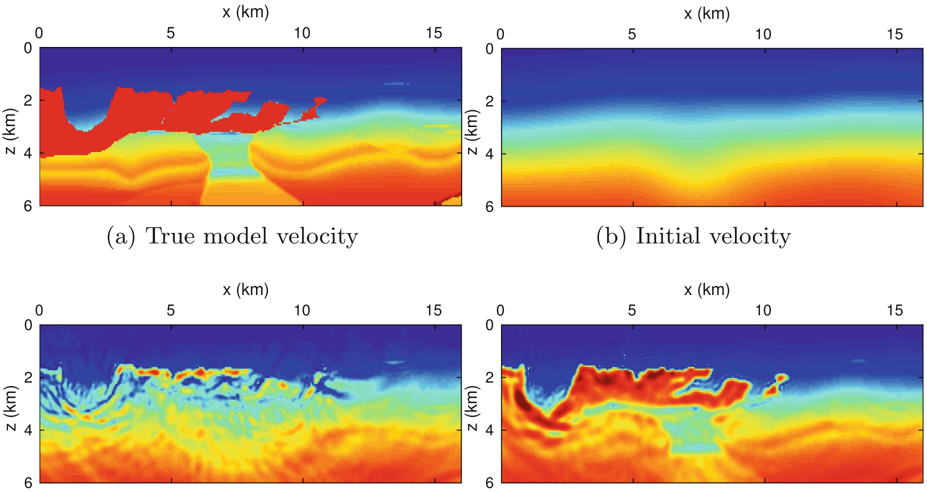
$$\frac{\partial J}{\partial m} = -\int_0^T u_{tt}(x,t)v(x,T-t) = -\int_0^T u(x,t)v_{tt}(x,T-t) \tag{11}$$

In the acoustic setting, the $v_{tt}(x,t)$ is equivalent to the wavefield with the second order time derivative of $\frac{\partial J}{\partial f}$ being the source. The change of the misfit function only impacts the source term of the back propagation, particularly the second order time derivative of $\frac{\partial J}{\partial f}$. For L^2 norm, the term is $2(f_{tt}(x,t)-g_{tt}(x,t))$, and for H^{-1} norm it becomes $2(g(x,t)-f(x,t))$. For trace-by-trace W_2 norm, the second order time derivative of $\frac{\partial W_2^2(f,g)}{\partial f}$ is $2\left(\frac{g(x,t')-f(x,t)}{g(x,t')}\right)$ where $t' = G^{-1}F(t)$, the optimal coupling of t for each trace.

Compared with L^2 norm, the source term of H^{-1} does not has the two time derivatives and therefore has more of a focus on the lower frequency part of the data. Lower frequency components normally provide a wider basin of attraction in optimization. The source term of W_2 is similar to the one of H^{-1} norm, but the order of signal g in time has changed with the optimal map for each trace at receiver x. The optimal couplings often change the location of the wavefront. For example, if g is a shift of f, then the wavefront of g will be mapped to the wavefront of f even if two wavefronts do not match in time. The change of time order in g also helps generate a better image under the reflectors when we back propagate the source and compute the gradient as in (11).

5 Numerical Example

In this section, we use a part of the BP 2004 benchmark velocity model [2] (Fig. 4a) and a highly smoothed initial model without the upper salt part (Fig. 4b) to do inversion with W_2 and L^2 norm respectively. A fixed-spread surface acquisition is used, involving 11 shots located every 1.6 Km on top. A Ricker wavelet centered on 5 Hz is used to generate the synthetic data with a bandpass filter only keeping 3 to 9 Hz components. We stopped the inversion after 300 L-BFGS iterations.



(a) True model velocity

(b) Initial velocity

(c) Inversion result using L^2

(d) Inversion result using W_2

Fig. 4. Large scale FWI example

Here we precondition the data with function $P(f) = a \cdot f + b$ to satisfy the nonnegativity and mass balance in optimal transport. Inversion with trace-by-trace W_2 norm successfully construct the shape of the salt bodies (Fig. 4d), while FWI with the conventional L^2 failed to recover boundaries of the salt bodies as shown by Fig. 4c.

6 Conclusion

In this paper, we revisited the quadratic Wasserstein metric from the optimal transport theory in the application of seismic inversion. The desirable properties of convexity and insensitivity to noise make it a promising alternative misfit function in FWI. We also analyze the conventional least-squares inversion (L^2 norm), the integral wavefields misfit function (H^{-1} norm) and the quadratic Wasserstein metric (W_2) in terms of the model parameter gradient using the adjoint-state method. The analysis further demonstrate the effectiveness of optimal transport ideas in dealing with cycle skipping.

Acknowledgments. We thank Sergey Fomel, Junzhe Sun and Zhiguang Xue for helpful discussions, and thank the sponsors of the Texas Consortium for Computational Seismology (TCCS) for financial support. This work was also partially supported by NSF DMS-1620396.

References

1. Alford, R., Kelly, K., Boore, D.M.: Accuracy of finite-difference modeling of the acoustic wave equation. Geophysics **39**(6), 834–842 (1974)
2. Billette, F., Brandsberg-Dahl, S.: The 2004 BP velocity benchmark. In: 67th EAGE Conference & Exhibition (2005)
3. Engquist, B., Froese, B.D.: Application of the Wasserstein metric to seismic signals. Commun. Math. Sci. **12**(5), 979–988 (2014)
4. Engquist, B., Froese, B.D., Yang, Y.: Optimal transport for seismic full waveform inversion. Commun. Math. Sci. **14**(8), 2309–2330 (2016)
5. Huang, G., Wang, H., Ren, H.: Two new gradient precondition schemes for full waveform inversion. arXiv preprint arXiv:1406.1864 (2014)
6. Liu, J., Chauris, H., Calandra, H.: The normalized integration method-an alternative to full waveform inversion? In: 25th Symposium on the Application of Geophpysics to Engineering & Environmental Problems (2012)
7. Métivier, L., Brossier, R., Mrigot, Q., Oudet, E., Virieux, J.: An optimal transport approach for seismic tomography: application to 3D full waveform inversion. Inverse Prob. **32**(11), 115008 (2016)
8. Plessix, R.E.: A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. Geophys. J. Int. **167**(2), 495–503 (2006)
9. Stolt, R.H., Weglein, A.B.: Seismic Imaging and Inversion: Volume 1: Application of Linear Inverse Theory, vol. 1. Cambridge University Press, Cambridge (2012)
10. Tarantola, A., Valette, B.: Generalized nonlinear inverse problems solved using the least squares criterion. Rev. Geophys. **20**(2), 219–232 (1982)
11. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
12. Virieux, J., Operto, S.: An overview of full-waveform inversion in exploration geophysics. Geophysics **74**(6), WCC1–WCC26 (2009)
13. Yang, Y., Engquist, B., Sun, J., Froese, B.D.: Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion. arXiv preprint arXiv:1612.05075 (2016)